DropConnected Neural Networks Trained on Time-Frequency and Inter-Beat Features for Classifying Heart Sounds

Edmund Kay, Anurag Agarwal

Engineering Department, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ , UK

E-mail: ek360@cam.ac.uk

23 March 2017

Abstract. Automatic heart sound analysis has the potential to improve the diagnosis of valvular heart diseases in the primary care phase, as well as in countries where there in neither the expertise nor the equipment to perform echocardiograms. An algorithm has been trained, on the PhysioNet open-access heart sounds database, to classify heart sounds as normal or abnormal. First, the heart sounds are segmented using an open-source algorithm based on a hidden semi-Markov model. Following this, the time-frequency behaviour of a single heartbeat is characterized by using a novel implementation of the continuous wavelet transform, mel-frequency cepstral coefficients, and certain complexity measures. These features help detect the presence of any murmurs. A number of other features are also extracted to characterise the inter-beat behaviour of the heart sounds, which helps to recognize diseases such as arrhythmia. The extracted features are normalized and their dimensionality is reduced using principal component analysis. They are then used as the input to a fully-connected, two-hidden-layer neural network, trained by error backpropagation, and regularized with DropConnect.

This algorithm achieved an accuracy of 85.2% on the test data, which placed third in the PhysioNet/Computing in Cardiology challenge (first place scored 86.0%). However, this is unrealistic of real-world performance, as the test data contained a dataset (dataset-e) in which normal and abnormal heart sounds were recorded with different stethoscopes. A 10-fold cross-validation study on the training data (excluding dataset-e) gives a mean score of 74.8%, which is a more realistic estimate of accuracy. With dataset-e excluded from training, the algorithm scored only 58.1% on the test data.

1. Introduction

In the USA, the prevalence of valvular heart disease in the population is 0.3% for 18-44 year olds, rising to 11.7% for those aged 75 and over [1]. This means valvular heart diseases are a significant public-health problem, whose diagnosis and treatment is important. An important first stage in the diagnosis of such diseases is auscultation,

where a doctor listens to the sounds generated by the heart through a stethoscope. Experienced practitioners can determine specific problems simply from the timing, intensity, and frequency of any heart murmurs [2]. However, current auscultation proficiency is poor and the percentage of correct diagnoses, by auscultation alone, is low. Mangione [3] studied the cardiac auscultation skills of trainee doctors from the USA, Canada, and the UK. He found that, on average, the trainees produced the correct diagnosis in 23% of cases, with a range of 0 to 58%.

If, on auscultation, the doctor hears an abnormal sound, the patient is referred for an echocardiogram, which is performed by a specialist and then analysed by a consultant to diagnose the valve disease. However, performing an echocardiogram is both expensive and time-consuming. Syed et al. [4] claim that around 80% of patients referred to cardiologists for echocardiograms have innocent heart murmurs and that this referral, in the USA, costs $300 to $1000 per patient. Shub [5] found that in the USA, between 1986 and 1989, the number of echocardiographic studies performed increased by 143%, costing $126 million. Despite the introduction of many more sophisticated diagnostic methods, such as echocardiograms and colour-flow Doppler techniques, Tavel [6] claims that cardiac auscultation still remains an important part of clinical medicine.

Since auscultation skills have declined and echocardiograms are both expensive and time-consuming to perform, there is a need for a fast, cheap method of producing accurate diagnoses of valvular heart diseases, especially in countries where there is neither the equipment nor the expertise to perform echocardiograms. Here, the possibility of automatic heart sound analysis is considered. In this system a recording of a patient's heart sounds, called a phonocardiogram (PCG), would be made via a stethoscope. Then the system would produce an automatic diagnosis of any valve disease present. This could reduce the number of missed diagnoses in the primary care phase as well as produce more accurate diagnoses in countries where echocardiograms cannot be performed.

There have been many previous attempts to diagnose heart diseases from PCGs. However, as noted by Liu et al. [7], these studies suffer from a number issues such as: not using separate test and training sets when evaluating the algorithm's performance; or using small, hand-picked datasets with little variety of pathologies. Also, each study used a different dataset, making it difficult to determine the relative performance of various approaches. To address these issues, an open-access database of heart sounds was compiled by Liu et al. [7]. This was then used in a machine learning challenge, run jointly by the Computing in Cardiology conference and the online resource for physiological data, PhysioNet. In this challenge 3,153 of the recordings were released for competitors to use as data to train their algorithms, while 1,277 recordings were kept hidden in order to evaluate the performance of each algorithm. Full details of the challenge are given by Clifford et al. [8].

A typical PCG classification system is described in figure 1. The main aim of this paper is to produce an algorithm which can differentiate between normal and abnormal PCGs. This paper gives an extended analysis of the algorithm originally described in

our conference paper [9].

Our approach mimics that used by doctors to diagnose heart diseases via auscultation. In traditional auscultation, doctors try to determine each individual heartbeat, using the S1 and S2 heart sounds. They then listen for any murmurs between S1 and S2, which can indicate pathology. They can also diagnose arrhythmia by listening to the timing of the S1 and S2 heart sounds [2]. Using this methodology as a blueprint, our algorithm first segments the heart sounds into S1, systole, S2, and diastole. Then, both the temporal and spectral content of the signal are extracted. Finally, we add features that describe the timing of S1 and S2. These features form the input to an artificial neural network, which learns to distinguish between normal and abnormal heart sounds.

2. Datasets

The training data in the open-access heart sounds database was obtained from a number of different sources, and these are labelled as datasets a – f in table 1. These were all recorded by different doctors using different stethoscopes and contain different numbers of normal and abnormal signals. Heart sounds were labelled as normal or abnormal by doctors using an echocardiogram (where available) as well as auscultation. Signal quality was assessed by database's compiler, based on whether they thought the signal was too noisy to realistically be classified. Full details of the datasets are given by Liu et al. [7]. The first thing of note is that dataset-e, which contains the majority of the

Dataset	Recording Modality	N (G)	N (P)	A (G)	A (P)
a	WAM ES	116	1	276	16
b	Litmann E4000 ES	295	91	73	31
c	Custom ES	7	0	20	4
d	Prototype ES	26	1	26	2
e	N: Microphone or PE sensor A: 3M Littman ES	1780	91	146	37
f	JABES digital ES	78	2	31	3
Total	N/A	2302	186	572	93

Table 1: Datasets in the open-access heart sounds database with their recording modality. WAM = Welch Allyn Meditron, ES = electronic stethoscope, PE = piezoelectric, N = normal recordings, A = abnormal recordings, G = good quality recordings, P = poor quality recordings.

PCGs, has its normal and abnormal PCGs recorded with different sensors. This means that it cannot reliably be used in any test/validation data, as one can not tell whether an algorithm trained on dataset-e has learnt to distinguish between normal and abnormal recordings or whether it has simply learnt to recognise between two different recordings

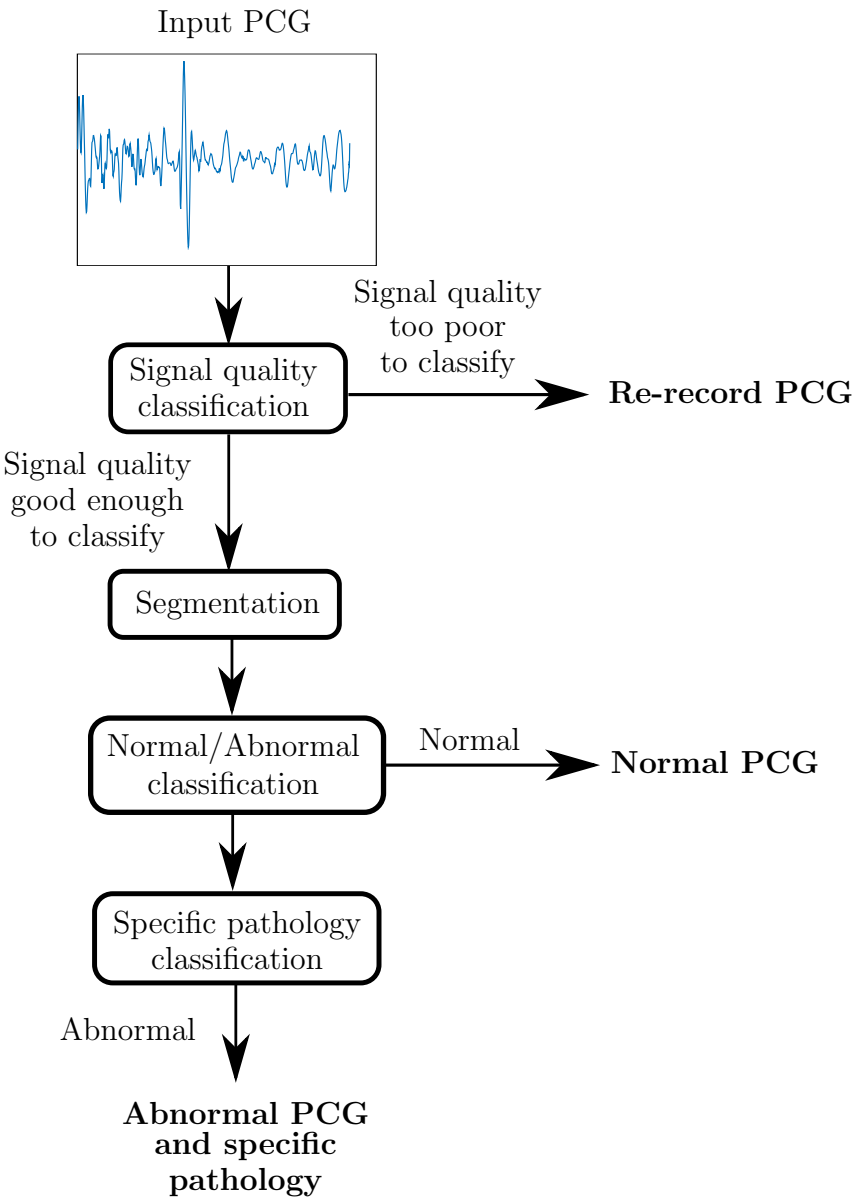


Figure 1: Schematic of an automatic auscultation system

modalities. However, dataset-e may be used in the training data to try and improve performance across the other datasets.

Also, it can be seen that each dataset is unbalanced. If each dataset was left as it is, the classifier would be more likely to diagnose as normal or abnormal based on which the majority class is in that dataset. If this heart sound analysis was done in the real-world then the prevalence of valvular heart disease, as well as the sensitivity-specificity trade-off for the given healthcare system, would determine the ratio of normal to abnormal signals in the training set. However, since this is currently unknown, then each dataset will be modified so that there are roughly equal numbers of normal and

abnormal recordings in each dataset. To produce a dataset called "Balanced Challenge" all the noisy signals are removed and the datasets are balanced as described in table 2. This will help give a better idea of the accuracy of the normal/abnormal classifier.

Dataset	N (G)	N (P)	A (G)	A (P)	Notes
a	232	0	276	0	Normals all repeated
b	73	0	73	0	1 in 4 normals kept at random
c	14	0	20	0	Normals all repeated
d	26	0	26	0	Nothing done
e	178	0	146	0	1 in 10 normals kept at random
f	39	0	31	0	1 in 2 normals kept at random
Total	562	0	572	0	

Table 2: Composition of the "Balanced Challenge" dataset. N = normal recordings, A = abnormal recordings, G = good quality recordings, P = poor quality recordings.

Finally, we note that if a PCG is classified as abnormal by the normal/abnormal classifier, then it should be possible to further classify the specific pathology of the murmur (the final step in figure 1). However, it is not possible to produce an algorithm to do this accurately with the current dataset. Table 3 shows each dataset from the open-access heart sounds database and the various pathologies present in the "abnormal" recordings. It shows that each pathology is only present in one dataset (except coronary artery disease which is present in datasets b and e). Therefore, if an algorithm was trained, on the whole database, to recognise specific pathologies, it would not be possible to tell if the algorithm was identifying different pathologies or simply recognising the different recording modalities in each dataset. In order to get more specific diagnoses of abnormal PCGs, a large database of different pathologies, all recorded with the same stethoscope, is required.

Dataset	Pathologies present (number of recordings)
a	MVP (137), Benign (118), AD (17), MPC (23)
b	CAD (151)
c	MR (17), AS (17)
d	No specific pathologies given (30)
e	CAD (335)
f	No specific pathologies given (33)

Table 3: Datasets in the open-access heart sounds database with the pathologies present in the abnormal recordings. MVP = mitral valve prolapse, AD = aortic disease, MPC = miscellaneous pathological conditions, CAD = coronary artery disease, MR = mitral regurgitation, AS = aortic stenosis.

3. Segmentation

To classify heart sounds as normal or abnormal, first an algorithm for segmenting heart sound recordings into S1, systole, S2, and diastole, is used. We used the segmentation algorithm supplied for the challenge, which was initially written by Schmidt et al. [10] and later improved by Springer et al. [11]. This algorithm extracts a variety of features which are then used to train a duration-dependent hidden semi-Markov model to label the PCG. The performance of this algorithm on each of the datasets is given in table 4 (summarised from Liu et al. [7]). This shows that segmentation algorithm, in general, works well but performs poorly on dataset c.

The reason is that it is difficult for the algorithm to segment signals containing murmurs that suppress S1 and S2 sounds [2]. Table 5 shows the proportion of recordings in datasets a–d that contain audible heart murmurs. Comparing tables 4 and 5, we see that datasets containing a high percentage of murmurs and noisy signals are more prone to segmentation inaccuracies on individual heartbeats.

Dataset	Recordings correctly segmented (%)	Beats correctly segmented (%)
a	71.1	88.4
b	67.1	74.1
c	32.3	58.5
d	43.6	80.5
e	86.4	90.3
f	64.9	83.1
All, no e	66.2	83.4
All	79.3	88.3

Table 4: Performance of the segmentation algorithm, no e = excluding dataset e. A recording is said to be correctly segmented if all heartbeats in that recording are correctly segmented

Dataset	Abnormal signals (%)	Signals with audible murmur (%)	Noisy signals (%)
a	71.4	20.1	4.2
b	26.4	10.0	24.9
c	77.4	64.5	12.9
d	50.9	27.7	5.5

Table 5: Percentage of signals with murmurs in datasets a–d

The poor segmentation performance on recordings in datasets c and d could propagate through to the later stages of the algorithm, creating an upper-bound on the performance of the normal/abnormal classifier. Specifically, if the algorithm is performing poorly on signals where there is a heart murmur, this could reduce the

sensitivity. It should be noted that it is possible to produce an accurate PCG classifier without segmentation [12]. However, our algorithm segments the PCG so that we can design features for the normal/abnormal classifier that are specific to our knowledge of murmurs and the cardiac cycle.

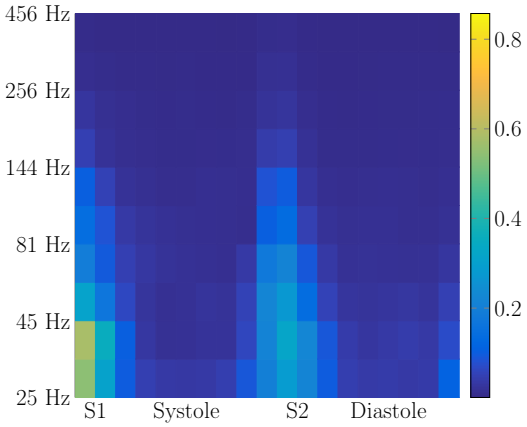
4. Feature Extraction

Once the heart sounds have been segmented, features are extracted to best represent the heart sounds to the classifier.

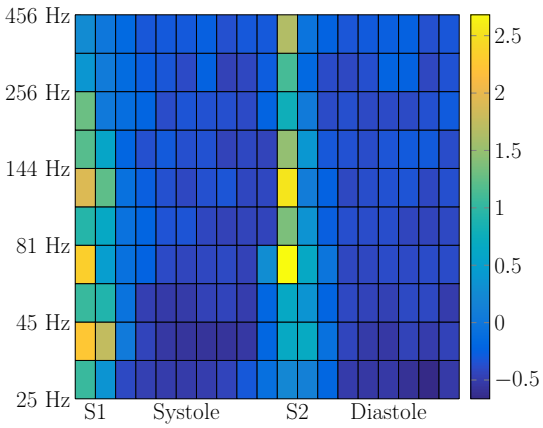
4.1. Wavelet Transform

By listening to a patient's heart, an experienced clinician can diagnose a wide range of pathologies by the timing and frequency of any murmurs present [2]. Therefore, we aim to produce a feature which shows the time and frequency behaviour over one cardiac cycle. This is done using the continuous wavelet transform (CWT), with the Morlet wavelet as the mother wavelet. The CWT is evaluated at 11 frequencies which are logarithmically spaced to give a better resolution at lower frequencies, where S1 and S2 sounds. Although this is at the expense of a worse resolution at high frequencies, the majority of heart murmurs produce broadband sounds which should show up in two to three frequency bins in the CWT (figure 3b). Increasing the number of frequency bins gives better resolution at the expense of increased complexity. The number of frequency bins was optimized using the best 10-fold cross-validation score on the training data.

The CWT is then normalized, at each frequency level, by subtracting by its mean and dividing by its standard deviation (across all time). This normalization helps to show up any murmurs present. Figure 3a shows the raw CWT for a typical beat in recording c0028 [7] (diagnosed as aortic stenosis by echocardiogram), in which a clear systolic murmur can be heard. Figure 3b shows the CWT for the same typical beat after normalization at each frequency. These figure show that the systolic murmur (between 150 and 250 Hz) can be seen much more clearly after normalization. Following this, the CWT is averaged into 20 time bins per heartbeat, 3 in S1, 7 in systole, 3 in S2, and 7 in diastole. This CWT is then averaged over heartbeats which are well correlated with each other. This is done because, while recording the sound signal, stethoscope movements can lead to varying amplitudes, resulting in some parts of the cardiac cycle being artificially louder than others. The beats which are well correlated are determined by finding the two beats with the minimum euclidean distance between their CWT coefficients at each frequency. Then any other beats within 50% of this minimum euclidean distance, from both of the two beats, are also averaged over. This leads to a time-frequency representation of a typical heartbeat which has 20 discrete points in time and 11 discrete points in frequency. A typical wavelet feature vector for a normal and abnormal recording are shown in figures 2b and 3b respectively. These figures show that, for the aortic stenosis patient (figure 3b), between 150 and 250 Hz,

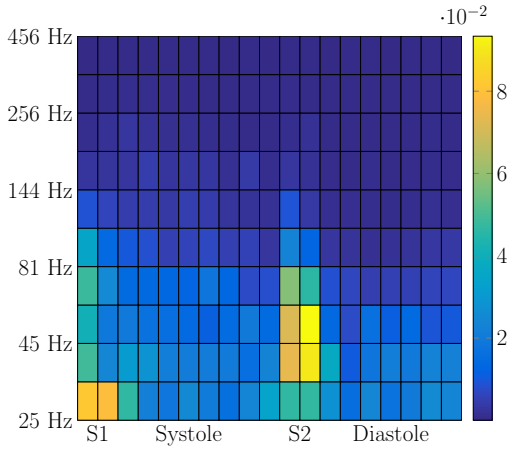


(a) Raw CWT for a typical beat

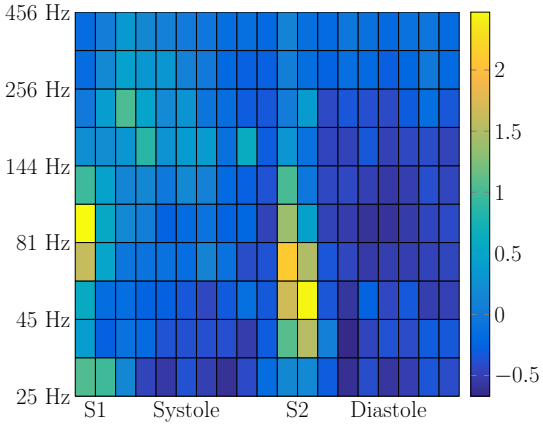


(b) CWT normalized at each frequency for a typical beat

Figure 2: A normal heart sound (c0011)



(a) Raw CWT for a typical beat



(b) CWT normalized at each frequency for a typical beat

Figure 3: An abnormal heart sound (c0028 - aortic stenosis)

the sound in the systolic phase has a higher amplitude relative to the fundamental heart sounds. This is as expected, as aortic stenosis has been shown to produce a broadband murmur in this frequency range [13].

4.2. Mel-frequency Cepstral Coefficients:

Mel-frequency cepstral coefficients (MFCCs) have been widely used in speech recognition [14], and have also been shown to be useful in heart sound classification [15]. In order to calculate the MFCCs, the signal is divided into the same 20 frames in time per heartbeat as for the CWT. Following this, a periodogram is found of each segment, with a 10% overlap either side. A Hamming window is used to reduce spectral leakage. The periodogram gives values for the signal power in 40 evenly spaced frequency bins

between 0 (not inclusive) and the Nyquist frequency (inclusive). As for the CWT, this periodogram is then averaged over beats which are well correlated with each other, at each frequency (see section 4.1 for full details). The periodogram is then filtered using a filter bank, in which frequencies, f, are equally spaced in the mel-scale (equation 1).

$$Mel = 1125 \log \left(1 + \frac{f}{700}\right). \tag{1}$$

The MFCCs are obtained by taking the logarithm of each of the filtered periodograms, and then taking a discrete cosine transform (equation 2) of each of the 20 frames.

$$MFCC(t,k) = \sum_{n=1}^{N} \log\left(P_{filt}(t,n)\right) \cos\left(\frac{k\pi}{N}\left(n - 0.5\right)\right) \tag{2}$$

$MFCC\left(t,k\right)$ gives the k^{th} cepstral feature of the t^{th} time frame. $P_{filt}(t,n)$ is the filtered power at time frame t for the n^{th} filter bank. N is the number of filter banks used.

Finally, the last cepstral feature is removed to give the final MFCC feature vector used for classification.

4.3. Inter-beat features

The features described in sections 4.1 and 4.2 help to determine the time-frequency characteristics of a typical beat in the PCG. However, these miss the differences in behaviour between different heartbeats in the cycle, which can indicate pathologies such as arrhythmia. Therefore, we add features that help to characterize this inter-beat behaviour.

The features, which were supplied for the challenge, were used [7]. These are the mean and standard deviation of: the length of one heart cycle; the length of S1; the length of systole; the length of S2; the length of diastole; the ratio of systolic length to whole heart cycle length; the ratio of diastolic length to whole heart cycle length; the ratio of systolic length to diastolic length; the ratio of mean systolic amplitude to mean S1 amplitude; and the ratio of mean diastolic amplitude to mean S2 amplitude.

4.4. Complexity

Features which characterise the complexity of the signal are also extracted. These have also been used by Schmidt et al. [16]. First, a periodogram is found with 20 time frames per heart cycle (3 in S1, 7 in systole, 3 in S2, 7 in diastole) and 5 equally spaced frequency frames. This is obtained in the same way as for the MFCCs (section 4.2) and is then normalized between 0 and 1. Following this, the spectral entropy (SE) is obtained as

$$SE\left(t\right) = -\sum_{f} P_{xx}\left(t,f\right) log\left[P_{xx}\left(t,f\right)\right]. \tag{3}$$

Also the unbiased standard deviation (SD), skewness (SK), and kurtosis (KT) of the power spectrum at each frequency, are obtained. It is found that these features only

improve the performance marginally, which is likely to be due to the fact the similar information is given in the features from sections 4.1 and 4.2.

The spectral entropy, standard deviation, skewness, and kurtosis make up 35 features which describe the complexity of the signal.

5. Feature Selection

Feature extraction results in a features vector of length 675. The make-up of the features vector is summarized in table 6. Features are then normalized by subtracting their

Feature	CWT	MFCC	Inter-beat	Complexity
Length	220	400	20	35

Table 6: Make-up of the features vector

means and dividing by their standard deviations (across the whole training set). After normalization, all features are subjected to a Student's t-test to determine whether they are significantly different between normal and abnormal recordings. Any feature with a test statistic less than the student's test statistic, from a two-tailed test at the 5% significance level, is removed. Then, for any pair of features which are highly correlated (a covariance greater than 0.9), one of them is removed (the one with the lowest t-statistic). Finally, a principal component analysis (PCA) is used to reduce the dimensionality of the features vector. The results in this paper are obtained by projecting the features vector onto its first 50 principal components.

6. Classification

The classification algorithm is based on a fully-connected, two-hidden-layer neural network, trained by error backpropagation [17]. The hyperbolic tangent activation function is used for all the neurons in the network except in the final layer, where the softmax activation function is used. The log-likelihood cost function is used. The hyper-parameters chosen for training the networks are given in table 7. In order to militate

Parameter	Value
Number of epochs	150
Mini-batch size	8
Learning rate	0.05
L2-regularization parameter	0
Momentum coefficient	0.3

Table 7: Parameters used for training the neural network

against overfitting, two types of regularization are used. The first is L2-regularization,

where a w^2 term is added to the cost function (where w is the weight along an individual neuron) to penalize large weights in the network. The second is DropConnect, which is described by Wan et al. [18]. For all the results here, the percentage of neurons which are randomly removed from each layer is 20% for the neurons between the input and the hidden layers, and 50% for the neurons between the two hidden layers and between the second hidden and the output layers. These values were found to give the right level of regularization and were optimized using the best 10-fold cross-validation score on the training data.

7. Results

The results of running the normal/abnormal classifier on a number of different datasets are shown in table 8. The scoring function used is the same as the one described by Liu et al. [7]. Table 8 shows that using the whole dataset and Springer's segmentation

Dataset	Hand, Score % (σ)	Springer, Score % (σ)
All challenge (stratified by dataset)	88.7 (2.3)	87.0 (2.3)
Balanced Challenge	81.2 (2.6)	79.1 (3.7)
Balanced Challenge, e removed from all	75.3 (4.1)	74.8 (5.1)
Balanced Challenge, e removed from test	73.3 (4.2)	71.8 (3.8)
Leave-one-out a (no e in training)	64.0	63.0
Leave-one-out b (no e in training)	66.4	64.5
Leave-one-out c (no e in training)	95.9	97.1
Leave-one-out d (no e in training)	58.7	57.3
Leave-one-out f (no e in training)	50.7	54.7
Leave-one-out a (e in training as balanced)	63.0	64.8
Leave-one-out b (e in training as balanced)	56.1	59.1
Leave-one-out c (e in training as balanced)	86.6	87.0
Leave-one-out d (e in training as balanced)	55.4	51.7
Leave-one-out f (e in training as balanced)	53.2	52.5

Table 8: Results for the normal/abnormal classifier. The first four rows are results from 10-fold cross-validations, with 10 repeats per fold. All the leave-one-out tests are the mean of 10 repeats. σ = variance. Hand = all data segmented using hand-segmented labels. Springer = all data segmented using Springer's algorithm [11].

algorithm [11], our algorithm is able to achieve a score of 87.0%. However, for the reasons discussed in section 2, this is unrealistic. An estimate of real-world performance is 74.8%, given by the row in which the dataset is "Balanced Challenge, e removed from all". This is a better estimate because the data does not contain the e-dataset and the number of normal and abnormal recordings are roughly equal. This real-world score of 75% is obtained from a sensitivity of 76% and a specificity of 74% (with the confusion

matrix given in table 9). Balancing the dataset has helped to ensure that sensitivity and specificity are roughly equal.

	Classified as Normal	Classified as Abnormal
Normal Signals	259	93
Abnormal Signals	102	324

Table 9: Confusion matrix showing results on validation data across 10 folds of cross-validation obtained on the dataset "Balanced Challenge, e removed from all". Note that the number of normal signals is lower than the sum of the normal signals in datasets a–d, and f. This is because dataset-a contains repeated normal signals. It was ensured that (for every fold of cross-validation) any repeated signals, which were common to both training and validation sets, were removed from the validation set.

The optimal ratio of sensitivity to specificity depends on the healthcare system in which this algorithm is being used. In a higher-income country with a well funded healthcare system, it may be beneficial to improve the sensitivity at the expense of specificity. This is because the healthcare system is more likely to accept the burden of increased referrals for echocardiogram in order to make sure that more people with heart disease are picked up in the primary care phase. However, in a developing country, the opposite may be true. With resources spread thinly, it may not be acceptable to refer patients without a disease for an echocardiogram and treatment. Therefore improving the specificity might be beneficial, even at the expense of reducing the sensitivity. It is possible to change the sensitivity to specificity ratio of the algorithm by changing the ratio of normal to abnormal signals in the training data.

Table 8 also shows that using the current algorithm, it is better to exclude the e-dataset completely than to try and include it in the training data (when it is not in the test data). Including the e-dataset worsens the performance on 2/2 10-fold cross validation studies (by an average of 2.5%) and on 8/10 leave-one-out tests (by an average of 4.3%). It is possible, however, that a dedicated transfer learning algorithm could make use of the e-dataset to improve performance.

Table 8 also shows that using hand-segmented PCGs (for both the training and validation data) improves the performance on 4/4 10-fold cross-validation studies (by and average of 1.6%), but makes it worse and on 6/10 leave-one-out tests (by an average of 0.17%). However, the results for the hand-segmented signal are unrealistic as real-world test signals would have to be segmented by an algorithm.

The fact that the 10-fold cross-validation tests are significantly better than the leave-one-out tests (excluding dataset c), suggests that the algorithm is very sensitive to the recording modality.

7.1. Challenge Results

When this algorithm was submitted to be evaluated on the test data, a number of different networks were trained with a range of hyper-parameters and different training sets. For example, a 5-fold cross-validation is done with networks trained on different hyper-parameters. The networks are then ensembled based on their score on the validation data and their diversity (measured by which recordings they incorrectly classified). Each of the networks in the ensemble classifies the heart sounds. The final classification is given by the majority.

The best result obtained by this ensemble of networks, on the test data, was 85.2%. This was the third best performance in the challenge, with 86.0% being the best score. This is, however, an over-estimation of real-world performance, since dataset-e was used in the test data (see section 2 for explanation).

The updated algorithm was also scored on the test set, but the overall performance was significantly worse (58%). This is because 69% of recordings in the test set are from dataset-e [7], and our algorithm is no longer trained on this dataset. Performance on unseen datasets g and i is also poor (table 10). This shows that the algorithm is sensitive to the recording type and struggles to generalize from one dataset to another.

Table 10 shows that the updated algorithm performs significantly better on datasets c and d. The slight reduction in the score on dataset-b does not necessarily mean that the algorithm is performing worse. Since 35.6% of recordings from dataset-b in the test set are labelled as noisy, any small changes in performance could be due to fortunate classification of the pathologies underlying noisy signals [7]. The performance of the normal/abnormal classifier on this dataset will be easier to determine when a signal quality classifier is implemented (as in figure 1).

Dataset	Original Algorithm	Updated Algorithm
b	74.7%	70.4%
c	77.5%	95.0%
d	58.3%	87.5%
e	93.6%	45.9%
g	57.3%	46.6%
i	50%	49.6%
All	85.2%	58.1%

Table 10: Results on different datasets in unseen test data

8. Conclusions and Future Work

An algorithm capable of classifying heart sounds as normal or abnormal has been developed. It starts by segmenting the heart sounds' recording into S1, systole, S2,

and diastole using an open-source algorithm, which is 88.3% accurate on all heartbeats in the database. Then, a total of 675 features are extracted from a recording. The time-frequency behaviour of the recording is characterized by using a continuous wavelet transform (CWT). The CWT is normalized at every frequency to clearly show any high frequency (150-450 Hz) murmurs. A key step to reduce unwanted noise and improve robustness is to perform an ensemble average over well-correlated beats. This also helps to militate against stethoscope movements that artificially change the recording. Similar techniques are used to get mel-frequency cepstral coefficients, and certain spectral complexity measures. All these features give a full picture of the time-frequency behaviour of a typical heartbeat in the recording, which helps detect the presence of any murmurs. Inter-beat features are added to look for differences between heartbeats in the recording, which helps to detect diseases such as arrhythmia.

In order to deal with the varying magnitudes of the features, they are normalized across the entire training data. To reduce overfitting, the dimensionality of the features vector is reduced by projecting it onto its first 50 principal components. Classification is done using a fully-connected, two-hidden-layer neural network, trained with error backpropagation and regularized using DropConnect.

This algorithm obtained an accuracy of 85.2% on the test data, which placed third in the PhysioNet/Computing in Cardiology challenge (first place scored 86.0%). However, this is unrealistic of real-world performance, as the test data contained a dataset (dataset-e) in which normal and abnormal heart sounds were recorded with different stethoscopes. A 10-fold cross-validation study on the training data (excluding dataset-e) gives a mean score of 74.8%, which is a more realistic estimate of accuracy.

Obtaining more specific diagnoses of abnormal heart sounds was considered. However, it was shown not to be possible with the current database. Specific diagnoses will require a large database of different pathologies, all recorded with the same stethoscope.

The current algorithm classifies any input signal as normal or abnormal. However, if a doctor makes a poor quality recording then it would be better to recognize this and tell them to re-record the signal. Therefore, an algorithm will be developed that is capable of recognizing a good quality recording from a poor quality one.

References

[1] V T Nkomo, J M Gardin, T N Skelton, J S Gottdiener, C G Scott, and M Enriquez-Sarano. Burden of valvular heart diseases: a population-based study. *The Lancet*, 368(9540):1005–1011, 2006.

[2] J Constant. *Essentials of Bedside Cardiology*. Humana Press Inc., New York, USA, 2003.

[3] S Mangione. Cardiac auscultatory skills of physicians-in-training: a comparison of three English-speaking countries. *The American Journal of Medicine*, 110(3):210–216, 2001.

[4] Z Syed, D Leeds, D Curtis, F Nesta, R A Levine, and J Guttag. A Framework for the Analysis of Acoustical Cardiac Signals. *IEEE Transactions on Biomedical Engineering*, 54(4):651–662, 2007.

[5] C Shub. Echocardiography or auscultation? How to evaluate systolic murmurs. *Canadian Family Physician*, 49:163–167, 2003.

[6] M E Tavel. Cardiac auscultation: a glorious past – and it does have a future! *Circulation*, 113(9):1255–1259, 2006.

[7] C Liu, D Springer, Q Li, B Moody, R A Juan, F J Chorro, F Castells, J M Roig, I Silva, A E W Johnson, Z Syed, S E Schmidt, C D Papadaniil, L Hadjileontiadis, H Naseri, A Moukadem, A Dieterlen, C Brandt, H Tang, M Samieinasab, M R Samieinasab, R Sameni, R G Mark, and G D Clifford. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(11):2181–2213, 2016.

[8] G D Clifford, C Liu, D Springer, , B Moody, Q Li, R Abad, J Millet, I Silva, A Johnson, and R G Mark. Classification of Normal/Abnormal Heart Sound Recordings: the PhysioNet/Computing in Cardiology Challenge 2016. *Proceedings of Computers in Cardiology*, page in press, 2016.

[9] E Kay and A Agarwal. DropConnected Neural Network Trained with Diverse Features for Classifying Heart Sounds. In *Proceedings of Computers in Cardiology*, Vancouver, Canada, 2016.

[10] S E Schmidt, C Holst-Hansen, C Graff, E Toft, and J J Struijk. Segmentation of heart sound recordings by a duration-dependent hidden Markov model. *Physiological Measurement*, 31(4):513–529, 2010.

[11] D Springer, L Tarassenko, and G Clifford. Logistic Regression-HSMM-based Heart Sound Segmentation. *IEEE Transactions on Biomedical Engineering*, 63(4):822–832, 2016.

[12] M Zabihi, A B Rad, S Kiranyaz, M Gabbouj, and A K Katsaggelos. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In *Proceedings of Computers in Cardiology*, Vancouver, Canada, 2016.

[13] D Kim and M E Tavel. Assessment of severity of aortic stenosis through time-frequency analysis of murmur. *CHEST*, 124(5):1638–1644, 2003.

[14] X Huang, A Acero, and H-W Hon. *Spoken Language Processing*. A Guide to Theory, Algorithm, and System Development. Prentice Hall, Upper Saddle River, New Jersey, USA, 2001.

[15] P Wang, C S Lim, S Chauhan, J Y A Foo, and V Anantharaman. Phonocardiographic Signal Analysis Method Using a Modified Hidden Markov Model. *Annals of Biomedical Engineering*, 35(3):367–374, 2006.

[16] S E Schmidt, C Holst-Hansen, J Hansen, E Toft, and J J Struijk. Acoustic Features for the Identification of Coronary Artery Disease. *IEEE Transactions on Biomedical Engineering*, 62(11):2611–2619, 2015.

[17] C M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK, 1995.

[18] L Wan, M Zeiler, S Zhang, Y L Cun, and R Fergus. Regularization of Neural Networks using DropConnect. In *International Conference on Machine Learning*, pages 1–9, 2013.